RUNNING HEAD: Balkanization of Probability

Balkanization and Unification of Probabilistic Inferences

Chong-Ho Yu, Ph.D.

Paper presented at American Educational Research Association

San Diego, CA (April, 2004)

Revised April 12, 2005

Chong-Ho Yu, Ph.D.

Psychometrician

Cisco systems/Aries Technology

PO Box 612

Tempe AZ 85248

USA

Email: chonghoyu@yahoo.com

Website: http://www.creative-wisdom.com

Abstract

Many research-related classes in social sciences present probability as a unified approach based upon mathematical axioms, but neglect the diversity of various probability theories and their associated philosophical assumptions. Although currently the dominant statistical and probabilistic approach is the Fisherian tradition, the use of Fisherian significance testing of the null hypothesis and its probabilistic inference has been an ongoing debate. This paper attempts to explore the richness and complexity of the ideas of probability with the emphasis on the relationships between Fisherian and other probability theories. First, it clarifies the differences between Fisher and Jeffreys and explains the background history relating to Fisher's quest for certainty. Second, it explains the differences between Fisher and Pearson and explains the limitations of the Fisherian approach. In addition, it argues that although Fisher criticized the Bayesian school for its alleged lack of objectivity, Fisher's quest for certainty is driven by his subjective faith in experimental methods, eugenics and Darwinism. Last, it will briefly introduce the synthesized approaches by Berger and Pawitan, respectively, as a possible remedy.

Balkanization and Unification of Probabilistic Inferences

Introduction

Use of hypothesis testing and its probabilistic inference has been an ongoing debate for over two decades (Harlow, Mulaik, & Steiger, 1997). While many authors (e.g. Hubbard & Bayarri, 2003) identified that the current form of hypothesis testing is a fusion of incompatible methodological traditions established by Fisher and Neyman/Pearson, respectively, very few people are aware of the historical background from which the dispute arose. When I was in graduate school, my mentor always advised me, "Be intimate with the data. Always try to understand how and where the data come from." By the same token, it is beneficial for researchers to know what the social and academic cultures were when Fisher developed his school of thought, what philosophy and worldview Fisher embraced, what research goals he tried to accomplish, and how Fisherian probability is related to other schools of thought. By knowing this information, we will be in a better position to judge the appropriateness of use of hypothesis testing and probabilistic inference based upon theoretical distributions. In other words, the inquiry of the meanings of statistical and probabilistic inferences can be illuminated by analysis in the perspectives of philosophy and history of science.

Although philosophers, such as Carnap (1950) and Hacking (1990, 2001), and historian of science Howie (2002) had devoted tremendous efforts to analyze statistics and probability with a wide horizon, cross-disciplinary dialogues are not common in this topic. For example, in the beginning of the 20th century, statistician Fisher and philosophers von Mises and Reichenbach independently devoted efforts to construct their own versions of frequency theories of probability. Frequency theories of probability are characterized by attaching probability to some "reference class" or "reference set." To be specific, if R is the reference class, n is the number of events in R, and m is the number of events in X, within R, then the probability of X, relative to W, is m/n. Although Fisher and von Mises/Reichenbach define the reference set differently, their theories are in the same vein in terms of conceptualization. However, Salmon (1967), a student of Reichenbach, credited Reichenbach as the developer of the frequency theory without a single

word about Fisher. In discussing philosophical foundations of probability theory, Weatherford (1982) also ignored Fisher entirely; he emphasized only the role of von Mises and Reichenbach in the development of frequency theory. Nonetheless, statisticians are equally self-centered. During the early 20th century, von Mises was not widely cited in statistics texts or debates of the Royal Statistical Society, in which Fisher played an active role (Howie, 2002). While giving guidelines to accessing probability, risk and statistics, Everitt (1999) mentioned the work of Fisher only. Not surprisingly, neither von Mises nor Reichenbach appears on Everitt's radar screen.

Owing to the Balkanization of probability theories, probability remains a confusing concept. For example, currently the Fisherian hypothesis testing school dominates quantitative methodology, but few people realize that the current hypothesis testing is a fusion of Fisher's and Neyman/Pearson's probability theories and statistical methodologies, which contain many incompatible elements. Even though the differences between Fisher and Neyman/Pearson was discussed by statisticians, the relationships among biometry, Mendelism, Darwinism, and Fisherism were rarely mentioned. In addition, the dispute between Fisher's frequency view of probability and the Bayesian view of probability is always simplified as a battle between an objective-oriented approach and a subjective-oriented one; nonetheless, Fisher's quest for objectivity and certainty is driven by his bias towards eugenics and evolution.

This paper carries multiple facets. First, it attempts to clarify the differences between Fisher and Jeffreys, and the background history relating to Fisher's quest for certainty. It argues that Fisher's quest for certainty is driven by his faith in experimental methods, eugenics and Darwinism. Second, it illustrates the differences between Fisher and Pearson, and why the so-called certain and objective approach is not really objective and certain. In spite of all these differences, unification of probabilistic inferences is still possible. Hence, at the end it will briefly introduce the synthesized approaches by Berger and Pawitan, respectively.

Differences between Fisher and Jeffreys

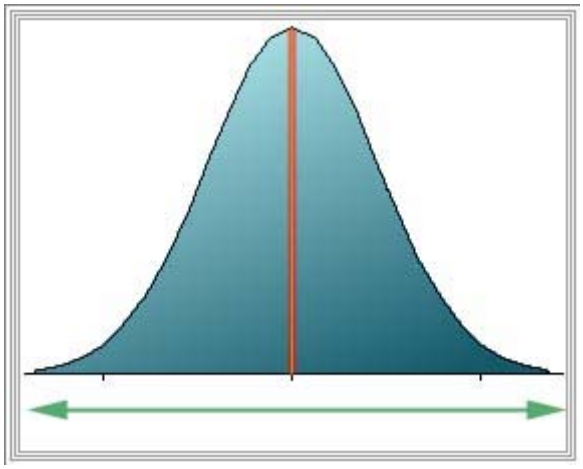
In Fisher's early career, he emphasized the certainty of science and proposed a testing model that yields a dichotomous answer. On the other hand, Harold Jeffreys embraced the Bayesian approach, which

defines probability in terms of the degree of belief. In the Bayesian approach, a researcher starts with a prior probability and the posterior probability is updated by subsequent evidence. According to Jeffreys, use of prior probabilities is no embarrassment, since priors represent the expertise or background knowledge a scientist brings to the data. If the scientist does not know much about the subject matter, even complete ignorance is a state of knowledge that could be quantified in probabilistic terms.

The issue of the hypothetical infinite population is another "battlefield" for Fisher and Jeffreys. In the Fisherian school, inference is a process of connecting the observed sample and the unobserved population using the sampling distribution as a bridge. Jeffreys asserted that scientists are not interested in the properties of some remote, hypothetical population. Rather, they are concerned with particular cases here and now. However, Fisher said nothing about how to transfer a probability statement from a population to an individual (Howie, 2002). Put simply: It doesn't help me at all that a doctor tells me the treatment is effective for the cancer patient population in the long run; I just want to know whether the treatment would work for me. Further, the target population to which inferences are made is infinite. To Jeffreys, Fisher's view of probability that involves infinities should be rejected as meaningless. Jeffreys maintained that researchers should be interested in the obtained data, not a sampling distribution averaged over all possible samples (Howie, 2002).

Fisher, who disliked vagueness and subjectivity, was strongly opposed to Jeffreys and his Bayesian colleagues (Howie, 2002). To Fisher it was absurd to confine probability to actual data, whose properties vary from time to time, from place to place, and from person to person; probability should carry objective and invariant properties that can be derived from mathematics. As a competent mathematician, Fisher constructed three criteria of desirable properties of estimators to the unknown population, namely, unbiasedness, consistency, and efficiency (Eliason, 1993). A detailed mathematical demonstration of these properties is beyond the scope of this paper; nevertheless, the following brief description of Fisher's approach demonstrates how Fisher elegantly constructed an objective approach to statistics and probability even if the hypothetical population is unknown in distribution and infinite is size.

Figure 1. Unbiased estimator.



In Figure 1, the bell-shaped curve denotes the hypothetical distribution. The red line represents the population parameter while the yellow line represents the estimation. If the estimated parameter is the same as the true parameter, this estimation is considered unbiased. Nonetheless, this Fisherian notion is not universally accepted for unbiasedness could be viewed as a property of a sampling distributon, not a property of an estimate. An estimator has variance or dispersion. In Figure 2 the green line with arrows at both ends indicates that the estimator may fall somewhere along the dispersion. An efficient estimator is the one that has achieved the lowest possible variance among all other estimators, and thus it is the most precise one. Moreover, the goodness of the estimation is also tied to the sample size. As the sample size increases, the difference between the estimated and the true parameters should be smaller and smaller. If this criterion is fulfilled, this estimator is said to be consistent. Hence, researchers can make probabilistic inferences to hypothetical populations using these objective criteria.

Fisher and experimental methods

At first glance, Fisher's "objective" approach sounds more convincing than his Bayesian counterpart, and it is no wonder that it is welcomed by most quantitative researchers. Undoubtedly the Fisherian hypothesis testing has currently overshadowed Bayesianism and other quantitative methodologies. As Berger (2000) noted, today many universities do not offer Bayesian-related courses in their statistical

training. Once a journal reviewer commented that my article should focus on mainstream quantitative methods, such as Fisherian hypothesis testing, rather than "marginal" approaches such as Bayesianism. On the other hand, Bayesianism is very popular among philosophers (e.g. van Fraassen, Brad Armendt). Why frequentism dominates statistics and Bayesianism is popular in philosophy is a very interesting topic for historians of science.

The difference between the Fisherian and the Bayesian schools could be understood through the perspective of Fisher's quest for certainty. Fisher conducted research in biometrics and agricultural sciences, in which data were collected from experimental methods. On the other hand, his Bayesian rival Jeffreys, who was in geophysics and astrophysics, did not enjoy the luxury of what Fisher had and was able to collect fragmented and ambiguous data only. It is obvious that geophysicists and astrophysicists cannot manipulate the solar system and the earth core for experimentation. Direct observations are also difficult. Hence, it is no wonder that Jeffreys interpreted probability as a degree of belief due to inherent limitations. However, with a strong experimental background, Fisher asserted that objective science must have an empirical starting point and yield an answer with a high degree of certainty (Howie, 2002).

Darwinism, Mendelism, and Biometrics

Unfortunately, while Fisher criticized the subjectivity of the Bayesian approach, he might not be aware that he was affected by his own subjectivity—his pre-determined agenda on biological philosophy and political/social policy. On one hand, Fisher was very critical of Jeffreys and Pearson. For instance, he regarded the Bayesian approach as "more a consequence of insufficient schooling than a definite wish to advocate the epistemic interpretation" (cited in Howie, 2002, p.122). During the dispute between Fisher and Pearson in the 1920s, "Fisher kept up a steady barrage and rarely missed a chance to either attack Pearson directly or snipe at his advocacy of Inverse Probability." Fisher boldly claimed that his method of estimating population parameters was efficient and sufficient, but Pearson's methods were inefficient, insufficient, or inconsistent (Howie, 2002, p.66)

On the other hand, Fisher was very forgiving to Gregor Mendel, the father of genetics, even though he proved that Mendel was dishonest in interpreting the results of his genetics experiments (Press &

Tanur, 2001; Fisher, 1936). Mendel established the notion that physical properties of species are subject to heredity. In accumulating evidence for his views, Mendel conducted a fertilization experiment in which he followed several generations of axial and terminal flowers to observe how specific genes were carried from one generation to another. On subsequent examination of the data using Chi-square tests of association, Fisher (1936) found that Mendel's results were so close to the predicted model that residuals of the size reported would be expected by chance less than once in 10,000 times if the model were true. In spite of this rebuttal, Fisher was surprisingly polite to Mendel. For example, in telling that Mendel omitted details, Fisher wrote, “Mendel was an experienced and successful teacher, and might well have adopted a style of presentation suitable for the lecture-room without feeling under any obligation to complete his story by unessential details” (p.119). While discussing how Mendel lied about his data, Fisher wrote, “He (Mendel) is taking excessive and unnecessary liberties with the facts” (p.120). To explain why Mendel was wrong about his data, Fisher wrote, “It remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected” (p.132).

Interestingly enough, Fisher treated Darwin in a similar manner, probably due to his strong Darwinian orientation. In the development of his randomization test for paired data, Fisher used Darwin’s data on the relative growth rates of cross- and self-fertilized corn to demonstrate the merits of this non-parametric procedure. Although Fisher criticized that Darwin did not randomize the group assignment, he did not criticize other aspects of Darwin’s experimental design. After carefully checking Darwin’s description of the experiment, Jacquez and Jacquez (2002) found that this experiment indeed did not use true paired comparisons. They argued that although the foundation of Fisher’s randomization test is justified, indeed the data do not meet the rigorous criteria for paired data. Jacquez and Jacquez regarded this matter as being of a historical interest without making further implications. However, one may wonder what criticisms Fisher would have made if Pearson were the one who made such experimental errors.

This double standard, different treatments to Pearson on one hand, to Mendel and Darwin on the other hand, might be due to the fact that Fisher had adopted the Mendelian genetic and the Darwinian

evolutionary models. To be specific, one of Fisher's career goals is to synthesize biostatistics, Mendelism, and Darwinism (Howie, 2002; Provine, 1971). In late 19th century, Charles Darwin proposed that natural selection, in terms of survival for the fittest, is a driving force of evolution. Francis Galton, a cousin of Darwin, was skeptical of the selection thesis. Galton discovered a statistical phenomenon called regression to the mean (or median), which is the precursor of regression analysis. According to regression to the mean, in a population whose general trait remains constant over a period of generations, each trait exhibits some small changes. However, this change does not go on forever and eventually the trait of offspring would approximate that of the ancestors. For example, although we expect that tall parents give birth to tall children, we will not see a super-race consisting of giants after ten generations, because the height of offspring from tall people would regress toward the mean height of the population. According to Darwinism, small improvement in a trait across generations and the natural selection of keeping this enhanced trait make evolution possible, but Galton argued that the regression effect counter-balance the selection effect (Gillham, 2001)..

One of the major questions of evolution is whether variation of a trait is inheritable. In late 19th century Mendel gave a definite answer by introducing an elementary form of genetic theory. Mendel's theory was forgotten for a long while but it was re-discovered by de Vries in 1900. In contrast to the original Darwin's position that evolution is a result of accumulated small changes of traits, biologists who supported Mendel's genetics suggested the otherwise: evolution is driven by mutation and thus evolution is discontinuous in nature. By the end of 19th century and early 20th century, two opposing schools of thoughts were developed, namely, biometricians and Mendelians. Although Galton rejected the idea of small changes in trait as an evolutionary force, he was credited as the pioneer of biometrics for his contribution of statistical methods to the topic of biological evolution. These diverse views of the evolutionary and genetic theories are germane to the development of probabilistic and statistical inferences, because the notions of the infinite population in the Fisherian school and description of individual datasets in the Pearsonian school could be traced back to their positions in biological sciences (Yu, 2005).

Different approaches taken by Fisher and Pearson in Mendelism

One of the most vocal figures in the biometrics camp is Karl Pearson, the inventor of Product Moment Correlation Coefficient. By computing the correlation coefficients of physical traits among relatives sampled from human populations, Pearson concluded that there is no evidence that the variance of height among humans could be explained by heredity, and thus the correlational studies contradicted the Mendelian scheme of inheritance.

Fisher bluntly rejected Pearson's assertion by re-interpreting Pearson's data. Based on the same data set collected by Pearson for denying Mendelism, Fisher demonstrated that the hypothesis of cumulative Mendelian factors seems to fit the data very well. By re-formulating statistical procedures and probabilistic inferences, Fisher concluded that heritable changes in the Mendelian sense could be very small and evolution in the Darwinian sense could be very slow, and these subtle differences could be detected by Fisher's version of biometrics.

Their clash came to a crescendo in 1918 when Pearson, who served as a reviewer of the journal of Royal Society, rejected a paper submitted by Fisher regarding Mendelism and Darwinism. Fisher blamed the rejection on the paper being sent to "a mathematician who knew no biology" (cited in Morrison, 2002). With regard to the dispute on Mendelism and biometrics, several scholars explained how Fisher and Pearson differed in various aspects. For example, Norton (1975) argued that as a positivist, Pearson downplayed the role of causation in research. Instead, correlation plays a more central role in Pearson's formulation of theory. To be specific, if variables A and B are correlated, it does not necessarily imply that A causes B or vice versa. In Pearson's view, the ultimate essence of biological knowledge is statistical and there is no room for causal factors. The goal of statistical knowledge is descriptive, but not explanation. However, Fisher, as the inventor of randomized experimental design, did not reject the possibility of causal inferences. In addition, Morrison (2002) pointed out that Pearson did not regard knowledge as absolute truth and hence the end product of statistical analysis is conceptual modeling which serves as a approximation to the phenomenon that we observed. This view is in a head to head collision with the Fisherian approach. Fisher were disinterested in individualistic information, but

asserted that biological inferences should be made with reference to indefinitely large number of Mendelism characteristics. On the other hand, Pearson accepted that using large but finite populations is a cornerstone of biometric methods, but rejected the notion of infinite populations.

Eugenics as a social fashion

Fisher's synthesis of Mendelism, Darwinism, and biometrics is tied to the fashion of eugenics, a variant of Mendelism, in the late 19th and early 20th centuries (Brenner-Golomb, 1993; Giegerenzer et al, 1989). During that period of time Westerners were highly interested in eugenics--applied genetics. Many research endeavors were devoted to explaining why Western civilizations were superior to others (e.g., research on intelligence) and how they could preserve their advanced civilizations. According to Darwinism, the fittest species are the strongest ones who could reproduce more descendants. This notion seems to fit the social atmosphere very well. To be explicit, Darwinism could rationalize the idea that the West is stronger and thus fitter; it has the "mandate destiny" before the nature has selected the superior. However, the dispute between biometricians and Mendelians, as well as the dissent voice of Karl Pearson, were considered a hindrance to the advance of Darwinism. Fisher's research provided an answer to a question that was seriously concerned by Western policy makers and scholars. Under the Mendelian-Darwinian-Biometrician synthesis, Fisher suggested that the only way to ensure improvement of the nation was to increase the reproduction of high-quality people (Brenner-Golomb, 1993). Thus, Fisher's insistence on certainty is partly motivated by his enthusiasm in promoting certain political/social policies and his faith in the Mendelian-Darwinian-Biometrician synthesis.

It is not the author's intention to discredit Fisher or downplay the Fisherian methodology by illustrating his obsessive quest for certainty and his agenda in Eugenics. Rather, this background information can help us to understand the limitations of the Fisherian school. For psychological and educational researchers, it is legitimate to ask whether a methodology aimed at achieving certainty in biology for partly serving specific political agenda (eugenics) is fully applicable to social sciences in general.

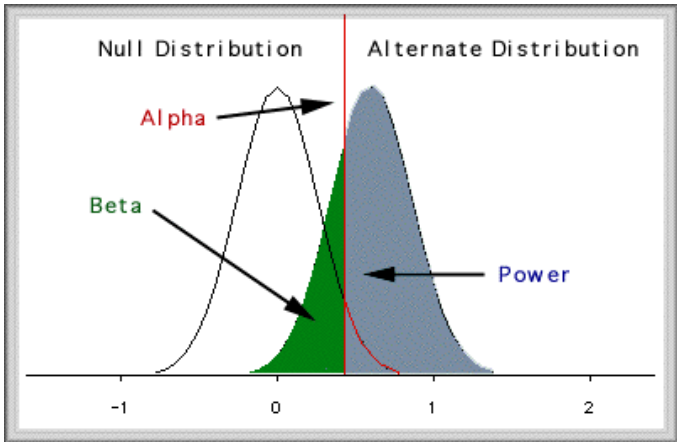
Difference between Fisher and Pearson in statistical testing and probabilistic inferences

It is important to keep in mind that under the synthesis of Mendelism, Darwinism, and Biometrics, the goals of Fisher was to develop a quantitative method, which is sensitive to the slowness of selection and the smallness of Mendelian changes, yet it could lead to a conclusion with a high degree of certainty or even a dichotomous answer, such as variation of species explained in genetics. With this background information the differences between R. A. Fisher and E. S. Pearson (the son of Karl Pearson) in statistical testing and probabilistic inferences will be more understandable.

The current form of hypothesis testing is a fusion of two schools of thought: Fisher and Neyman/Pearson (Lehmann, 1993). When Fisher introduced the null hypothesis (i.e., there is no difference between the control group and the treatment group), followers of this strategy tended to see that the only possible options are whether to reject the null hypothesis or not. Put simply, the conclusion is an either/or answer. To Pearson, testing a single hypothesis that only yields a simple and dichotomous answer is inadequate. Later Pearson introduced the concept of alternate hypothesis (i.e., there is a difference between the control group and the treatment group). However, the alternate hypothesis is unknown and thereby could be anything (e.g. a very huge difference, a large difference, a medium difference, a small difference, a very small difference, etc.). With the presence of alternatives, the conclusion is no longer dichotomous.

Further differences between the two schools can be found in the use of cut-off Alpha level. While Fisher advocated .05 as the standard cut-off Alpha level, Pearson (1933) did not recommend a standard level but suggested instead that researchers look for a balance between Type I and Type II errors. Statistical power is also taken into consideration for computing probabilities and statistics. Type I, Type II errors, Alpha, and power will be discussed in the section entitled “Neyman/Pearson model.”

Figure 2. Fusion of Fisher and Pearson models



Fisherian model

In Figure 2, the y-axis is the probability and the x-axis is the standardized score with the mean as zero and the standard deviation as one. The curve on the left hand side is the null distribution introduced by Fisher. It is important to note that this is the sampling distribution, which appears *in theory only*. It is derived from neither the population nor the sample. In theory, if there is no difference between the control and treatment groups in the population, the subtraction result is zero. However, there are always some sampling fluctuations due to measurement errors and other factors. In a thought experiment, if many samples are drawn from the same population, the difference is not exactly zero all the time. On some occasions it is above zero, and in some cases it is below zero. According to the Central Limit Theorem, when these scores are plotted, a bell-shaped distribution is formed regardless of the shape of the population distribution (Yu, Anthony, & Behrens, 1995). In the Fisherian methodology, a pre-determined Alpha level (red line) is set to guide the researcher in making a judgment about the observed sample. After the statistical attributes of the observed sample are found, the sample is compared against this theoretical sampling distribution. If the sample is located in the right hand side of the Alpha level, the data are said to be extremely rare, and thus the null hypothesis is rejected. Therefore, the region in the right hand side of the Alpha level is called the “region of rejection.”

Pearson model

Pearson enriched the methodology by introducing the concepts of alternate hypothesis, power, Type I and Type II errors (Beta). According to Pearson, it is not helpful to conclude that either there is no difference or some difference. If the null hypothesis is false, what is the alternative? Then development of Pearson's notion of alternate distributions may be partly tied to his father's disagreement with Galton on the nature of biological data. Galton believed that all biological data are normally distributed and variations should be confined within certain parameters. As mentioned before, Galton believed in regression to the mean, in which every naturally occurring variable has a fixed mean and all values of the variable should tend to scatter around the mean. On the other hand, Karl Pearson held a more open-ended view of distributions—the world should have more than one type of distribution. Data could take on a variety of shape, which could be skewed, asymmetrical, flat, J-shaped, U-shaped, and many others (Magnello, 1996).

Besides providing an alternate hypothesis, E. S. Pearson also changed the concept of probability from static and single-faceted to dynamic and multi-faceted. If the difference between the control and the treatment groups is small, it is possible that the researcher is unable to detect the difference when indeed the null hypothesis is false. This is called a Type II error, also known as "Beta." On the contrary, the researcher may also reject the null hypothesis when in fact there is no difference. In this case, the researcher makes a Type I error, also known as "false claim."

In the frequentist approach of E. S. Pearson, the validity of a test procedure is tied to the Type I error rate; a valid test should give an actual Type I error rate equal to the claimed Type I error rate. If the chosen Alpha cut-off is 0.05 but indeed the actual Type I error rate is 0.5, then the proclaimed conclusion is ten times more likely to be erroneous. Today tremendous research endeavors have been committed to Monte Carlo simulations and the development of other procedures for controlling the inflation of the Type I error rate. Behrens and Yu (2003) mocked it as the "neurosis of Type I error." They argued that error detection should focus on the anomaly of the data.

Under the frequentist logic of Pearson, several other probability concepts are introduced: Power, which is associated with the alternate hypothesis, is the probability that the null hypothesis is correctly rejected, whereas Beta is the probability of Type II error. In this dynamic model, power is a function of sample size, Alpha level, and the supposed mean difference between the two groups, which is also known as the "effect size." This configuration provides researchers a more versatile tool to conduct experiments and interpret probabilities. For more information on the relationships among null distribution, alternate distribution, power, Beta, Alpha level, effect size, and sample size, please view an animated illustration developed by Yu and Behrens (1995): http://www.creative-wisdom.com/multimedia/power.html

Philosophical shortcomings

Dichotomous character. There are several philosophical shortcomings in this integrated probability model. One of the problems is that the statistical result yielded from a testing is interpreted as a dichotomous answer: Either accept the hypothesis or reject the hypothesis. However, a dichotomous answer contradicts the very definition of probabilistic inference, which indicates uncertainty. In an attempt to amend this problem, Pearson (1955) admitted that the terms "acceptance" and "rejection" in statistical conclusions, which carry a connotation of absolute certainty, were unfortunately chosen. Rao's (1992) assessment of Fisher's work is helpful to clarify several misconceptions of dichotomous decisions in statistical testing:

> The decision (reject/not reject the null) is based on the logical disjunction … Such a prescription was, perhaps, necessary at a time when statistical concepts were not fully understood and the exact level of significance attained by a test statistic could not be calculated due to the lack of computational power…Fisher gives a limited role to tests of significance in statistical inference, only useful in situations where alternative hypotheses are not specified…Fisher's emphasis on testing of null hypotheses in his earlier writings has probably misled the statistical practitioners in the interpretation of significance tests in research work (p.46)

Rao is entirely correct in his assessment of Fisher's work. In his later career, Fisher started to realize the weaknesses of his methodology. First, he disapproved of the use of any standard Alpha level, though

he once supported it. He wrote, “No scientific worker has a fixed level of significance from year to year, and in all circumstances, he rejects hypothesis; he rather gives his mind to each particular case in the light of his evidence and ideas” (cited in Upton, 1992, p. 397). However, Fisher’s statement was a prescription about what scientists ought to do, not a description of what scientists did. Second, Fisher was opposed to handing over the judgment about whether or not to accept a hypothesis to an automated test procedure (cited in Mulaik, Raju, & Harshman, 1997, pp.78-79). Further, Fisher (1956) emphasized that the purpose of research is to gain a better understanding of the experimental material and of the problem it presents.

Unfortunately, up to the present day most researchers still adopt a conventional Alpha cut-off and run statistical tests in a mechanical manner; most researchers could not distinguish Fisher’s early view on probability and statistics from his later view. As a counter-measure, today some researchers de-emphasize the dichotomous character of hypothesis testing by asserting that the proper language of concluding a hypothesis testing should be “failed to reject the hypothesis” rather than “accepting the hypothesis” or “proving the hypothesis.” (Cohen, 1990; Parkhurst, 1985, 1990) In a similar vein, Lehmann (1993) gave researchers several practical suggestions:

> Should this (the reporting of the conclusions of the analysis) consist merely of a statement of significance or nonsignificance at a given level, or should a p value be reported? The original reason for fixed, standardized levels—unavailability of more detailed tables—no longer applies, and in any case reporting the p value provides more information. On the other hand, definite decisions or conclusions are often required. Additionally, in view of the enormously widespread use of testing at many different levels of sophistication, some statisticians (and journal editors) see an advantage in standardization; fortunately, this is a case where you can have your cake and eat it too. One should routinely report the p value and, where desired, combine this with a statement on significance at any stated level (p.1247).

Theoretical reference class. Further, determining the proper reference class is problematic. As illustrated in Figure 1, the sampling distribution is conceptualized through a *thought experiment*. We did not repeatedly sample the population. How could we know which distribution (reference class) should be

used to compare against the data? In this case, when an inference from sample to a population is made, strong assumptions must be imposed on the population (Yu, 2002).

Degree of confirmation. Another major problem with the Fisherian school, in Carnap (1950)'s term, is that the statistical result does not add anything to *the degree of confirmation*. First, in theory the sampling distribution is defined *in the long run*; the aim of hypothesis testing is to find out the probability that the sample is observed in the long run given that the null hypothesis is true. The controversy is: What is the long run? In theory the so-called "long run" could be infinite. Philosophers can easily cite the Humean challenge that nothing is conclusive in the long run because events occurring in the future may not resemble those in the past. Second, researchers are interested in knowing whether the theory is right given the evidence (data). However, comparing the data with the reference class can only tell us, given that the hypothesis is true, how likely we are to obtain the observed sample (evidence) in the long run. This is contrary to what other researchers want to know. In this case, it is doubtful whether the statistical result could add anything to the degree of confirmation.

Strong assumptions. On one hand it is true that the Central Limit Theorem secures the normality of the sampling distribution, no matter if the population has a skewed or a normal distribution. On the other hand, however, in many statistical tests the sample that we observed must still confirm to the normality assumption in order to make a valid inference. However, in reality data always departs from normality to certain degree. In this case, the inferential link between the sample, the theoretical sampling distribution, and the population is indeed very weak.

It is interesting that on one hand Fisher considered his probability theory as objective, but on the other hand, choosing remedies when parametric assumptions such as normality are violated is subjective. For example, when extreme scores affect the normality, Winsor suggested pulling the outliers toward the center because he believed that all observed distributions are Gaussian (normal) in the middle. Some other statisticians objected to the Winsorizing approach and recommended assigning different weights to outliers based upon their distance from the center. Nonetheless, Cliff was opposed to weighting because he insisted on the principle of one observation, one vote (Yu, 2002). The central question is: How should

we treat the sample in order to establish an inferential link among the sample, the sampling distribution, and the population? No matter how mathematical the winsorizing and weighting methods are, obviously they are not derived from self-evidential axioms.

Synthesis

With the increasing doubt of use of Fisherian/Neyman/Pearson hypothesis testing, several alternatives, such as effect size, exploratory data analysis, Bayesianism, and resampling methods have been proposed. Besides exploring alternate methodologies, "reforming" the Fisherian approach by blending divergent views of theories of probability inferences may be another viable alternative. In the past integrating Mendelism, Darwinsim, and Biometrics, as well as fusing Fisher, Neyman, and Pearson theories demonstrated that synthesis is possible even though apparently certain models seem to be incompatible. To be explicit, to researchers the question of methodology may not be an "either-or" question.

Substantive efforts have been devoted to attempts to remediate the Balkanization of probability. For example, Berger (2001) boldly synthesized Fisher, Jeffreys and Neyman's methodologies into a unified approach. In the Bayesian school, probability is conditional, while the concept "conditioning" is virtually absent from the frequentist school. Nevertheless, Berger extracted components from the three schools to formulate "conditional frequentist testing." However, what Berger did is methodological integration rather than philosophical synthesis. Berger admitted that his work was "motivated by the view that professional agreement on statistical philosophy is not on the immediate horizon, but this should not stop us from agreeing on methodology." (p.4) Philosophers may find this a-philosophical orientation unacceptable. Moreover, as a Bayesian, Berger (2000) asserted that the synthesis must be based upon the Bayesian theme because probability and statistics are about measuring uncertainty; the frequentist approach is useful to objectify the Bayesian estimation. Needless to say, this synthesis may not be welcomed by frequentists.

Pawitan (2000, 2001) also attempted to synthesize the frequentist and the Bayesian approaches. Unlike Berger, Pawitan placed the emphasis on the Fisherian school. Although Pawitan also viewed

probability as a measure of uncertainty, he accepted a “ladder of uncertainty,” a Fisherian idea introdcued in his last book *Statistical methods and scientific inference* (1956): Whenever possible, the researcher should base inference on probability statements, otherwise, it should be based on the likelihood. With the ladder of uncertainty as the foundation, Pawitan proposed the likelihood approach: Uncertainty can be expressed by both likelihood and probability, where likelihood is a weaker measure of uncertainty and probability allows objective verification in terms of long term frequencies. Pawitan argued that the likeihood approach is a compromise between Bayesianism and frequentism because this approach carries features from both factions.

Further, Pawitan developed the empirical likelihood approach by merging the likelihood and the bootstrap, which is a resampling method. In bootstrap, the sample is duplicated many times and treated as a virtual population. Then samples are drawn from this virtual population to construct an empirical sampling distribution. Like randomization exact test, the rationale of bootstrap is to counteract the theoretical aspect of the classical Fisherian approach by introducig empirical elements into the inference process. As I have shown, the classical Fisherian approach imposes several assumptions, such as normality and equal variances, on the sample. In reality, researchers always obtain “messy” data. Pawitan argued that the empirical likelihood approach could handle irregular data better than the classical one.

However, it seems that like Berger, Pawitan de-emphasized the role of philosophy. He argued that with the advent of Monte Carlo simulations performed in high-power computers, the Bayesian approach “can now be justified almost by the utilitarian principle alone, rather than by the orthodox philosophical reasons.” (p.5)

Discussion

Probability is a complicated and confusing concept. Not only did Fisher, Jeffreys, Pearson not agree with each other, but also in the earlier and later parts of his career, Fisher had different opinions on the same issues, such as use of a pre-determined Alpha level,. The synthesis between Fisher and Pearson amended several shortcomings in the Fisherian probability model, such as the introduction of Type I error,

Type II error, and power. Nonetheless, some researchers are doubtful that over-emphasis on reducing inflated Type I errors is justified.

Although Fisher employed rigorous mathematics to defend his view of probability in terms of hypothetical distributions and promoted his approach as objective science as opposed to subjective Bayesianism, his orientation is driven by his faith in Darwinism and eugenics. Indeed, various forceful assertions on statistics and probability theories made by Winsor, Cliff, and many others seem to be nothing more than professional opinions. As Berger said, agreement on statistical philosophy is not on the immediate horizon; current integration occurs on the methodological/computational level only. Thus, dialogues and collaborations between philosophers and statisticians are essential to the synthesis on the philosophical level.

Acknowledgments

Special thanks to Professor Brad Arment for his comments on this paper and Samantha Waselus for her professional editing.

References

Berger, J. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. Journal of American Statistical Association, 95, 1269-1276.

Berger, J. (2001 August). Could Fisher, Jeffreys, and Neyman have agreed on testing? Paper presented at the Joint Statistical Meetings, Atlanta, GA.

Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer, (Eds.). Handbook of psychology Volume 2: Research methods in Psychology (pp. 33-64). New Jersey: John Wiley & Sons, Inc.

Brenner-Golomb, N. (1993). R. A. Fisher's philosophical approach to inductive inference. In Keren G. & Lewis, C. (Eds.), A handbook for data analysis in the behavioral sciences (pp. 283-307). Hillsdale, NJ: LEA.

Carnap, R. (1946). Remarks on induction and truth. Philosophical and Phenomenological Research, 6, 590-602.

Carnap, R. (1950). Logical foundations of probability. Chicago, IL: University of Chicago Press.

Carnap, R. (1956). Meaning and necessity: A study in semantics and modal logic. Chicago, IL: University of Chicago Press.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Eliason, S. R. (1993). Maximum likelihood estimation: Logic and practice. Newbury Park: Sage.

Everitt, B. S. (1999). Chance rule: An informal guide to probability, risk, and statistics. New York: Springer.

Fisher, R. A. (1930). Inverse probability. Proceedings of the Cambridge Philosophical Society, 26, 528-535.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? Annals of Science, 1, 115-117.

Fisher, R. A. (1956). Statistical methods and scientific inference. Edinburgh: Oliver and Boyd.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Kruger, L. (1989). The empire of chance: How probability changed science and everyday life. Cambridge: Cambridge University Presss.

Gillham, N. W. (2001). A life of Sir Francis Galton: From African exploration to the birth of eugenics. Oxford: Oxford University Press.

Hacking, I. (1990). In praise of the diversity of probabilities. Statistical Science, 5, 450-454.

Hacking, I. (2002). An introduction to probability and inductive logic. Cambridge, UK: Cambridge University Press.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Mahwah, NJ: LEA.

Howie, D. (2002). Interpreting probability: Controversies and developments in the early twentieth century. Cambridge, UK: Cambridge University Press.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. American Statistician, 57, 171-178.

Jacquez, J. A., & Jacquez, G. M. (2002). Fisher's randomization test and Darwin's data -- A footnote to the history of statistics. Mathematical Biosciences, 180, 23-28.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, 88, 1242-1249.

Magnello, M. E. (1996). Karl Pearson's mathematization of inheritance: from ancestral hereidity to Mendelian genetics (1895-1909). Annals of Science, 55, 33-94.

Morrison, M. (2002). Modelling populations: Pearson and Fisher on Mendelism and Biometry. British Journal of Philosophy of Science, 53, 39-68.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A, 231, 289-337.

Parkhurst, D. (1990). Statistical hypothesis tests and statistical power in pure and applied science. In G. M. von Furstenberg (Ed.), Acting under uncertainty: Multidisciplinary conceptions (pp. 181-201). Boston, MA: Kluwer Academic Publishers.

Parkhurst, D.F. (1985). Interpreting failure to reject a null hypothesis. Bulletin of the Ecological Society of America, 66, 301-302.

Pawitan, Y. (2000). Likelihood: Consenus and controversies. Paper presented at the Conference of Applied Statistics in Ireland.

Pawitan, Y. (2001). In all likelihood: Statistical modeling and inference using likelihood. New York: Oxford University Press.

Pearson, E. S. (1955). Statistical concepts in their relation to reality. Journal of the Royal Statistical Society, Series B, 17, 204-207.

Press, S. J., & Tanur, J. M. (2001). The subjectivity of scientists and the Bayesian approach. New York: John Wiley & Sons.

Province, W. (1971). The origins of theoretical population genetics. Chicago, IL: The University of Chicago Press.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. Statistical Science, 7, 34-48.

Reichenbach, H. (1938). Experience and prediction: An analysis of the foundations and the structure of knowledge. Chicago, IL: University of Chicago Press.

Rodgers, J. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. Multivariate Behavioral Research, 34, 441-458.

Salmon, W. C. (1967). The foundations of scientific inference. Pittsburgh: University of Pittsburgh Press.

Upton, G. (1992). Fisher's exact test. Journal of the Royal Statistical Society. Series A (Statistics in Society), 155, 395-402.

von Mises, R. (1928/1957). Probability, statistics, and truth. London: The Macmillan Company.

von Mises, R. (1964). Mathematical theory of probability and statistics. New York: Academic Press.

Yu, C. H., & Behrens, J. T. (1995). Identification of misconceptions concerning statistical power with dynamic graphics as a remedial tool. Proceedings of 1994 American Statistical Association Convention. Alexandria, VA: ASA.

Yu, C. H., Anthony, S., & Behrens, J. T. (1995, April). Identification of misconceptions in learning central limit theorem and evaluation of computer-based instruction as a remedial tool. Paper presented at the Annual Meeting of American Educational Researcher Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 395 989)

Yu, C. H. (2002). An overview of remedial tools for violations of parametric test assumptions in the SAS system. Proceedings of 2002 Western Users of SAS Software Conference, 172-178.

Yu, C. H. (2005). History of science and statistical education: Examples from Fisherian and Pearsonian schools. 2004 Proceedings of the American Statistical Association, Statistical Education Section [CD-ROM], Alexandria, VA: American Statistical Association